Vision Sensory Substitution to Aid the Blind in Reading and Object Recognition

Pranab Gajanan Bhat, Deepak Kumar Rout, Badri Narayan Subudhi, and T. Veerakumar.
Department of Electronics and Communication Engineering, National Institute of Technology Goa, Ponda, India
pannubhat@gmail.com, deepak.y.rout@gmail.com, subudhi.badri@gmail.com, tveerakumar@yahoo.co.in

***Abstract*— Vision Sensory Substitution is a non-invasive rehabilitating procedure for the blind, which provides visual information to them via any of their functioning senses. In this paper, the authors propose a more engaging, interactive and integrated technology which will enable the visually impaired 'to read' the text and 'to see' their surroundings with the aid of computer vision. For reading of the text, Optical Character Recognition engine is utilized so that the blind can listen to the text which was intended to be read. For seeing objects and captioning them, features are extracted using Histogram of Oriented Gradients and the classification and labelling of the object via its features is done using K-Nearest Neighbor classifier. Results show that a high accuracies of 92% and 88% have been achieved for text extraction and caption generation respectively. The text and captions (also in text form) are finally converted into audio signals using text to speech converters.**

Keywords—Sensory substitution, Computer vision, Optical Character Recognition, edge detection, HOG, KNN

I. INTRODUCTION

Blindness describes the clinical condition whereby individuals completely lack the perception of light owing to sheer vision loss. They have to predominantly rely on their functioning senses except vision. Visual impairment is a vision loss condition characterized by the loss of visual functions at an organ level. According to a survey conducted by the World Health Organization (WHO) in 2014, there are an estimated 285 million people who suffer from blindness and visual impairments and it is also found that 90% of the visually impaired live in low-income settings [1].

Recent developments have led to numerous inventions making the visual world accessible to the blind and visually impaired. The blind lose their peripheral sensory system (the retina), but retain central visual mechanisms, thus still retaining their capacity to see [2]. Implants and stents are among various invasive techniques adopted by medical experts stimulating the surviving retinal cells, optic nerve or cortex. These invasive techniques have surgically associated risks, require extensive training for stimulating new neural connections, if necessary. These factors make invasive methods an expensive affair which is accessible only to the impaired who are financially better off. Alternately, non-invasive methods integrate human-centered computing and signal processing techniques in order to best harness the flexibility of a person's brain. These methods show promise in making visual information accessible to all sections of the blind and visually impaired [3]. It also proves advantageous in terms of relatively low cost and no health risks. Sensory substitution is the method of processing information attributed to a person's impaired modality by way of his or her unimpaired modality. In vision sensory substitution, the impaired modality is that of sight and the substituting modality can be auditory or touch. Vision sensory substitution blends machine learning, computer vision and image processing techniques for stimulating the unimpaired modality [4]. Here, the blind or visually impaired person is considered as a central processor and a camera is used to provide visual input to a signal processor which converts visual information into tactile or auditory information. The visual information extracted can be text or object in an image that needs to be processed and presented in terms of speech or touch to the visually impaired person.

Some of the vision sensory substitution technologies are briefly discussed here. Brain Port [5] aids the blind and visually impaired to navigate by acquiring visual data from a camera and processing it into vibrations felt on the tongue. VOICE vision technology, developed by Blue Edge Bulgaria, offers visual input through image to sound renderings for the blind [6]. Images acquired from the camera are scanned from left onwards to right, further being converted to sound by associating brightness with loudness and elevation with pitch. GuideCane's functionality is same as that of a white cane which serves the purpose of detecting presence of objects which the blind fail to see, but need to avoid walking into. The add-on in this technology is ultrasound sensors fitted on GuideCane which are capable of detecting obstacles in a wide sector in front of the user. This sensors aid to stay clear of the obstacles and is followed by the user which successfully manoeuvres through densely packed obstacles [7]. Schauerte et al. [8] have proposed a method for obstacle avoidance using Conditional Random Field (CRF) for aiding the blind. Some existing technologies like EyeMusic [9], visotoner [10], Google glass and Zungle glasses also show some potential but cannot be improved to specifically suit these needs of the blind and visually impaired. Hoang et al [11] have proposed a system to detect obstacles using a Kinect sensor. They then analyze the Kinect data to build the obstacle model without recognition, to warn the blind from running into the obstacle. Maidenbaum [12] et al. have used EyeMusic Sensory Substitution Device (SSD) for sonifying the on-screen content and conveying this sound to the blind for object detection.

978-1-5090-6734-3/17/$31.00 ©2017 IEEE

It can be summarized from the existing literature that there are different SSDs available for obstacle detection and avoidance for aiding the blind people for navigation. However, the detection as well as recognition of objects using SSD is still in the developing phase. As per author's knowledge, this is a novel attempt to integrate both text extraction and object recognition methods to aid the blind without the use of any existing SSD.

Here in this paper, we opt to develop a sensory substitution technology wherein the substitute modality is audio and propose a system which will actually help a blind person read the text and identify objects in a form that they can comprehend. Our work will enable the blind to listen (i.e. to read) to the text and furthermore, it will help them identify objects in the field of view of camera. Reading of the text is achieved by acquiring images of the text, followed by its extraction and eventually converting it to speech. For identifying the object, the edge map of the scene is obtained using the Canny edge detector. Features extracted from this edge map using Histogram of Oriented Gradients (HOG) are classified using K-Nearest Neighbor Classifier (KNN). Based on the classification, appropriate captions for the respective classes are generated tagging the object present in it.

The organization of this paper is as follows. In Section II, proposed method for implementing text to speech conversion and caption generation is narrated with the help of a block diagram and its results and discussions are presented in Section III. Section IV lists the conclusions.

II. PROPOSED METHOD

In this paper, we have proposed a system which integrates two utilities, amongst many, which the blind are deprived of. The first one allows them to read the text which they are unable to visualize due to their imparity. The second allows them to recognize the objects or even obstacles present in their surroundings. Initially, there is a need to discriminate between text and object in the field of view of the camera as it is continuously acquiring images. The authors address this problem by converting the color image to gray scale image and Histograms. Generally text is printed on a paper or sign board, are cases where both text and page/boards have a uniform color throughout. Hence the histogram of such image will be bimodal in nature. As regards to object in the surrounding, histogram is multimodal due to the presence of multi-color components present in the image. Hence, if the histogram is bimodal then it is considered as text and if it is multimodal then it is considered as an object in the surrounding. Fig. 1 shows the block diagram implementation of the discriminatory approach used.

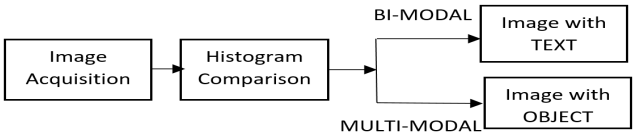


Fig. 1: Block diagram representation for discriminating text and object.

A. Reading of Text

One of the breakthroughs for enabling the blind people to read is development of Optical Character Recognition (OCR) technology [13], which extracts the text captured in an image. OCR efficiently extracts text information from an image via a three step process: Segmentation, Feature Extraction and Classification. Segmentation ensures that each character which is present in the image is correctly recognized. For feature extraction, usually template matching is deployed wherein the pixel matrix for each segmented character is used for feature extraction. By employing a correlation function *R*, the features of the segmented character are matched with the features of the template which are stored in the database. The character which receives the highest score after correlation is selected as the best match for that segmented character. The formulation of the correlation function R can be given as,

$$R = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \bar{A})^2\right)\left(\left(\sum_m \sum_n (B_{mn} - \bar{B})^2\right)\right)}} \quad (1)$$

In eq (1), A_{mn} denotes the segmented character of size m x n, and B_{mn} denotes the template stored in the database with which the character is being matched. The values denoted by $\bar{A}$ and $\bar{B}$ represent the average values of the segmented character and the template, respectively. The character is then classified as a certain alphabet with which it has the maximum correlation score. Once the text has been extracted it can be converted to audio with numerous text to speech converters available. This audio signals can enable the visually impaired 'to read' the text present in the image [14].Once the information content has been extracted this content is then converted to speech using text to speech synthesizer. A brief overview of implementing this part is shown in block diagram indicated in Fig. 2. The input here is the image having text obtained after the discrimination step performed previously.

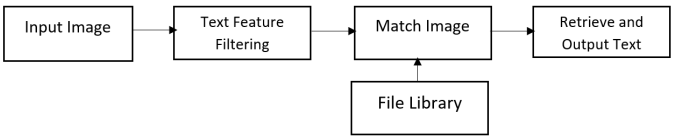


Fig. 2: Block diagram representation for converting text from image to speech.

B. Object Identification and Caption Generation

The discrimination stage indicates the presence of an object in the image. Then it is needed to identify the object and generate a suitable caption tagging it. This involves two crucial steps: content planning and surface realization [15]. Fig. 3 gives a brief overview in the form of a block diagram for the caption generation procedure.

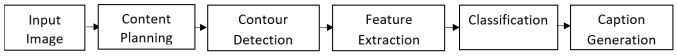


Fig. 3: Block diagram for generating captions for object present in the image.

Content Planning:

Content Planning acquires visual data from the image containing objects and a conditional random field (CRF) is used for predicting the best label for the image. Nodes of CRF correspond to the various contents of the image such as objects, various attributes responsible for appearance of the objects in given form and prepositions referring to a spatial relationship among the object-object pairs like above, below, over etc.. For an input image, a large set of object detectors across the image are run and only the high scores of detections are collected. The energy function, supposed to be minimized, for an image labelling (L) of image (I) will be,

$$E(L;I,T) = \sum_{i \in objs} F_i + \frac{2}{N-1} \sum_{i,j \in objPairs} G_{ij} \quad (2)$$

In eq (2), N is the number of objects in the image and 2/(N-1) is a factor which normalizes the contribution from object pair terms so that they contribute equally with the single object terms to the energy function. F_i and G_{ij} are detector score values for a single object and object pairs respectively. Formulation of F_i is and G_{ij} is given as,

$$F_i = \alpha_0\beta_0\varphi(obj_i, objDet) + \alpha_0\beta_1\varphi(attr_i, attrCl) + \alpha_1\gamma_0\varphi(attr_i, obj_i; text\,\mathrm{Pr}) \quad (3)$$

$$G_{ij} = \alpha_0\beta_2\varphi(prep_{ij}, prepFuns) + \gamma_1\varphi(obj_i, prep_{ij}, obj_j; text\ \mathrm{Pr}) \quad (4)$$

The function in the first, second and third term of eq (3) represents the detector scores for objects, attribution classification scores and prepositional relationship scores all by the trained data respectively. The terms in eq (4) describe the score for the prepositions to be put for the objects determining their locations with respect to other objects. α represents the tradeoff between image and text-based potentials. β represents the weighting between image-based potentials and γ represents the weighting between text-based potentials.

Contour Detection:

In order to mark the object within an image, we exploit the pixel size of the detected potential objects. The pixel size of different objects is compared with that of the entire image and a threshold is set. This is to segment the object of interest in the entire image. Appropriate pre-processing procedures like Gaussian filtering is performed. Once an object is detected, significant contours of it are retained using the canny edge detector. Since most of the information is contained in the edges and edges preserve the shape of the object, we deploy edge detection in our method.

Feature Extraction:

Once the contour is obtained, the next step is to extract the important features relating to this object. In this case the Histogram of Oriented Gradients (HoGs) descriptors have been used as the method for feature extraction. This method counts the number of occurrences of gradient orientations for each portion of the image. These portions into which image is further divided into smaller components are called as cells. For computing the HOG descriptors, the first step is to compute the gradient values. This is commonly done by filtering the image with the horizontal and vertical kernels which are given as,

$$D_x = [-1\ \ 0\ \ 1] \quad (5)$$

$$D_y = [1\ \ 0\ \ -1]^T \quad (6)$$

In this manner the derivatives of the image along the x and y directions are given by Ix and Iy. Once the derivatives have been obtained, the magnitude and orientation of the gradient associated are expressed as,

$$|G| = \sqrt{I_x^{\,2} + I_y^{\,2}} \quad (7)$$

$$\theta = \arctan\left(\frac{I_y}{I_x}\right) \quad (8)$$

The HoG descriptors are thus obtained for each cell, thereby giving an idea into the gradient direction for the pixels within the cell. This involves the creation of cell histograms. Based on the computed gradient values, each pixel in a cell casts a weighted vote. Each pixel casts a vote for orientation based on the closest bin the range from either 0 to 180 degrees or 0 to 360 degrees. The vote weight by each pixel can be calculated as either square of the magnitude of square root of the magnitude. In order to counter the changes created in the values due to pixel illumination and contrast, normalization is done. Normalization includes computing the values related to a larger portion of the image which is obtained by combining many cells. Thus, the HoG for the entire image can hence be obtained by combining the individual HoGs associated with each cell. These HoGs can further be normalized by calculating the measure of intensity associated with a larger section of the image called a block. Using the values thus generated for these blocks, the values for each cell within the given block are normalized. Once the values for each portion of the image is normalized, the final values denote the HoG feature values associated with the given pixel which are then stored so that further the objects can be classified based on these values.

Classification:

The classification of features into corresponding classes is performed by utilizing the KNN classifier. The procedure involved in KNN is finding minimum error between the features in the training class and test class. Since minimum distance also implies minimum error, we opt to compute the simplest Euclidean Distance for KNN. Since the training data set under consideration is expansive, the occurrence of error is reduced in our case. Once the object is classified using the KNN classifier,

it implies that the object has been identified and an appropriate caption has to be generated for this object.

Surface Realization and Caption Generation:

Surface realization is used for generating suitable captions for the identified object. It aims at generating natural language descriptions for the identified objects. This labelling encodes three kinds of information: objects present in the image (nouns), visual attributes of those objects (modifiers), and spatial relationships between objects (prepositions), if any. Some of the methodologies for surface realization are decoding using language models and Integer Linear Programming (ILP) model. The former uses Markov assumption wherein N^{th} word is dependent only on $(N-1)^{th}$ word as given in eq (9) whereas the latter uses language models and appropriate linguistic constraints to generate sentences for images.

$$P\left(x_i \middle| x_1,\ldots x_{i-1}\right)=P(x_i \middle| x_{i-N+1},\ldots x_{i-1}) \tag{9}$$

In our case, we have considered general sentence templates which can be chosen at random for a particular object. These templates are already stored in the memory. The identified description label is mapped from content planning into any one of the sentence templates based on the classifier output. Finally, the entire generated sentence is displayed for a particular object on the console.

Eventually, both the retrieved text from the OCR engine and the text generated from captions of the object are fed to the speaker of the earphones using a text to speech converter. Thus, the blind can listen to the text and also know about the objects present in their surroundings.

III. RESULTS AND DISCUSSIONS

The results obtained for text extraction using OCR tool and for generating caption for object are presented in this section.

A. Text extraction using OCR technology.

An RGB image as shown in Fig. 4(a) is the input acquired through the camera. After preprocessing the input image, it was binarized by appropriate thresholding. Fig. 4(b) and Fig. 4(c) represent the grayscale and binary images, respectively. This binary image is the input to the OCR engine which then extracts the text contained in the image. This text is printed on the console of Java as shown in Fig. 4(d). The text to speech conversion was performed for 50 images containing text of about 10,000 words. The measure of performance of this character recognition process is the accuracy defined as,

$$Accuracy\ (\%)=\frac{No.\ of\ correctly\ identified\ characters}{Total\ no.\ of\ characters} \tag{10}$$

The accuracy obtained for this OCR system was 92%. In cases of images which contained characters which were overlapping or when the image was scanned poorly, the characters were not correctly recognized.

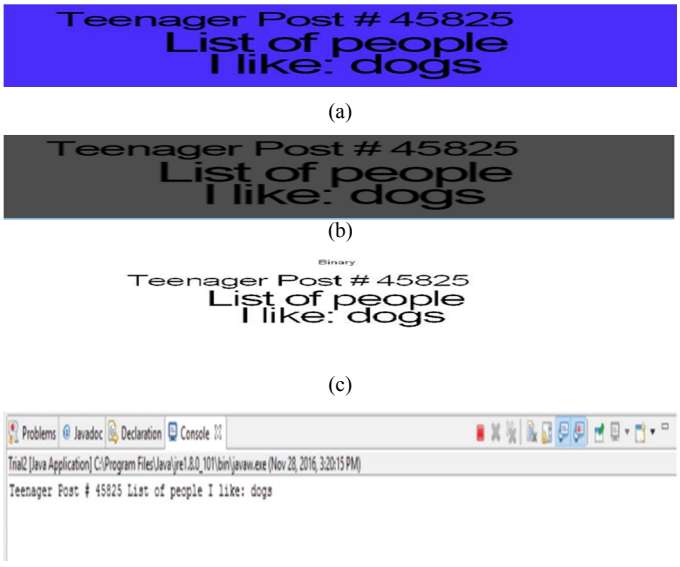


Fig. 4: Text extraction from image (a) Input RGB image, (b) grayscale image, (c) Binary image (d) Extracted text

B. Caption Generation

The object and its contours are detected using the pixel size thresholding and canny edge detector, respectively. The features of this contour are extracted using HOG. Post feature extraction, the image is classified into a particular class using the KNN algorithm. Object Detection was done for a total of ten classes, namely: aeroplane, car, crow, dog, elephant, guitar, laptop, shoe, stop sign and traffic cone. Each of these classes were denoted by classes from 1-10 respectively, wherein if an image was mapped to class 1 it was classified as an aeroplane. If it was mapped to class 2 it was classified as a crow and so on. The input images, their respective edges, HOG features and the caption generated are as shown in Figs. 5 (a)-(d), Figs. 6 (a)-(d), Figs. 7 (a)-(d), Figs. 8 (a)-(d) and Figs. 9 (a)-(d) for aeroplane, crow, standing dog, sitting dog and stop sign, respectively.

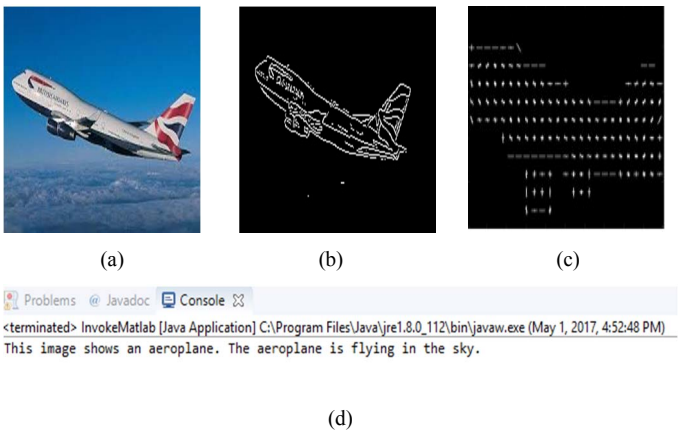
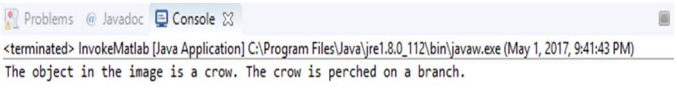


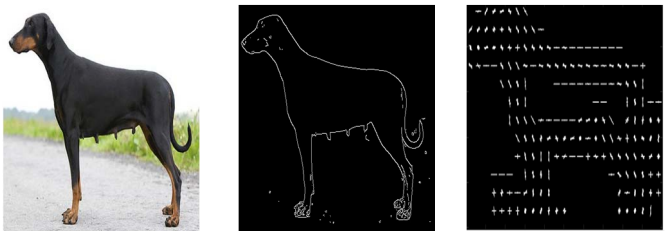
Fig. 5: Caption generation for an aeroplane (a) RGB Input (b) Canny edge detection (c) HOG features (d) Caption generated

(a) (b) (c)

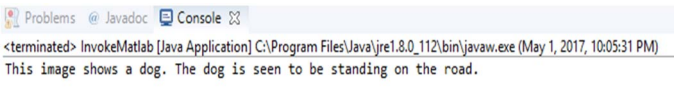


(d)

Fig. 6: Caption generation for a crow (a) RGB Input (b) Canny edge detection (c) HOG features (d) Caption generated



(a) (b) (c)



(d)

Fig. 7: Caption generation for a standing dog (a) RGB Input (b) Canny edge detection (c) HOG features (d) Caption generated

(a) (b) (c)

(d)

Fig. 8: Caption generation for a sitting dog (a) RGB Input (b) Canny edge detection (c) HOG features (d) Caption generated

(a) (b) (c)

(d)

Fig. 9: Caption generation for a stop sign (a) RGB Input (b) Canny edge detection (c) HOG features (d) Caption generated

The classification is gauged by accuracy which is defined as,

$$Accuracy\ (\%) = \frac{No.\ of\ correctly\ classified\ examples}{Total\ no.\ of\ examples} \quad (11)$$

Thus, the classification and caption generation for all ten classes consisting of 1,150 testing images was performed. The classification accuracy rate of 88% was obtained. The average time for generation of captions after acquiring the image is 0.6 seconds for all classes. In certain cases, an image of an elephant was wrongly recognized as that of a dog. In addition, some images of the class crow were incorrectly identified as aeroplane. These were a few of the cases due to which the accuracy of the system was reduced. Also in certain cases a sitting dog was wrongly identified as a standing dog, but was attributed to the same class of dog. The mistaken images can be attributed to the diverse test data set, excess background noise leading to KNN mistakenly associating a particular class to its nearest neighbor.

IV. CONCLUSIONS

We have proposed a vision to auditory sensory technology that follows a new direction as compared to conventional technologies. The OCR and Text-to-speech synthesizer were implemented to extract the text information from images to convert it to sound output. While considering images, the objects contained in the images were detected. Following which appropriate features were extracted in order to move on to the classification stage. Classification of images was done using the KNN algorithm and appropriate labels were assigned to the test data. Upon surface realization, the generated captions were converted to speech.

The proposed method works efficiently for recognition of only a single object present in the scene. It can be extended further for detecting multiple objects. For surface realization, we have generated captions using pre-stored captions from the memory. Captions generation can further be optimized by the use of natural language processing (NLP). NLP conjures up a sentence from scratch generating appropriate verbs and prepositions also

taking care of their placement in the sentence. The discrimination method using the histogram to discriminate text and object can be optimized further. Also, images taken from a camera can be used to provide real-time test data. The camera can be mounted on a spectacle frame worn by blind and visually impaired. The vision sensory technology could run on a portable processor that the person can use at any point in time.

REFERENCES

[1] WHO Media Centre, Visual impairment and blindness (2014). Retreived from http://www.who.int/mediacentre/factsheets/fs282/en/

[2] P. Bach-y-Rita and S. W. Kercel, "Sensory substitution and the human – machine interface,"*Trends in Cognitive Sciences*, vol. 7, no.12 ,pp. 541-546, Dec 2003.

[3] D. J. Brown and M. J. Proulx, "Audio–Vision Substitution for Blind Individuals: Addressing Human Information Processing Capacity Limitations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 5, pp. 924-931, Aug. 2016.

[4] S. Maidenbaum, R. Arbel, G. Buchs, S. Shapira and A. Amedi, "Vision through other senses: Practical use of Sensory Substitution devices as assistive technology for visual rehabilitation," *22nd Mediterranean Conference on Control and Automation,* Palermo, pp. 182-187, 2014.

[5] S. Upson, "Tongue Vision," *IEEE Spectrum,* vol. 44, no. 1, pp. 44-45, Jan. 2007.

[6] W. D. Jones, "Sight for sore ears [visual aids for the blind]," *IEEE Spectrum*, vol. 41, no. 2, pp. 13-14, Feb. 2004.

[7] S. Shoval, I. Ulrich and J. Borenstein, "NavBelt and the Guide-Cane [obstacle-avoidance systems for the blind and visually impaired]," *IEEE Robotics & Automation Magazine*, vol. 10, no. 1, pp. 9-20, Mar 2003.

[8] B. Schauerte, D. Koester, M. Martinez, R. Stiefelhagen, "Way to Go! Detecting Open Areas Ahead of a Walking Person," *European Conference on Computer Vision* , pp. 349-360, 2014.

[9] Abboud S, Hanassy S, Levy-Tzedek S, Maidenbaum S, Amedi A, "EyeMusic: Introducing a "visual" colorful experience for the blind using auditory sensory substitution," *Restor Neurol Neurosci*, pp. 247-257, 2014.

[10] M. Capp and P. Picton, "The optophone: an electronic blind aid," *Engineering Science and Education Journal*, vol. 9, no. 3, pp. 137-143, Jun 2000.

[11] V. Hoang, T. Nguyen, T. Le, T. Tran,T. Vuong, N. Vuillerme, "Obstacle detection and warning system for visually impaired people based on electrode matrix and mobile Kinect," *Vietnam Journal of Computer Science*, Vol. 4, Issue 2, pp. 71-83, May 2017.

[12] S. Maidenbaum, S. Abboud, G. Buchs, A. Amedi, "Blind in a Virtual World: Using Sensory Substitution for Generically Increasing the Accessibility of Graphical Virtual Environments," IEEE Virtual Reality (VR), 2015.

[13] R. Neto, N. Fonseca "Camera Reading For Blind People," *Procedia Technology,* vol. 16, pp. 1200-1209, 2014.

[14] C. Liambas and M. Saratzidis, "Autonomous OCR dictating system for blind people," *Global Humanitarian Technology Conference* , USA, 2016.

[15] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Lee, Y. Choi, A. Berg, T. Berg, "BabyTalk: Understanding and Generating Simple Image Descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891-2903, Dec. 2013.

[16] A. Singh and S. Desai, "Optical character recognition using template matching and back propagation algorithm," *International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, pp. 1-6, 2016.

[17] Y.Feng and M. Lapata, "Automatic Caption Generation for News Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 797-812, April 2013.